An effective fuel level data cleaning and repairing method for vehicle monitor platform

Article (Accepted Version)

An Effective Fuel Level Data Cleaning And Repairing Method for Vehicle Monitor Platform

Daxin Tian, *Senior Member, IEEE,* Yukai Zhu, Xuting Duan, Junjie Hu, Zhengguo Sheng, Min Chen, *Senior Member, IEEE,* Jian Wang, Yunpeng Wang

***Abstract*—With energy scarcity and environmental pollution becoming increasingly serious, the accurate estimation of fuel consumption of vehicles has been important in vehicle management and transportation planning towards a sustainable green transition. Fuel consumption is calculated by fuel level data collected from high precision fuel level sensors. However, in the vehicle monitor platform, there are many types of error in the data collection and transmission processes, such as the noise, interference, and collision errors are common in the high speed and dynamic vehicle environment. In this paper, an effective method for cleaning and repairing the fuel level data is proposed, which adopts the threshold to acquire abnormal fuel data, the time quantum to identify abnormal data, and linear interpolation based algorithm to correct data errors. Specifically, a modified Gaussian Mixture Model (GMM) based on the synchronous iteration method is proposed to acquire the thresholds, which uses the Particle Swarm Optimization (PSO) algorithm and the steepest descent algorithm to optimize the parameters of GMM. The experiment results based on the fuel level data of vehicles collected over one month prove the modified GMM is superior to GMM-EM on fuel level data, and the proposed method is effective for cleaning and repairing outliers of fuel level data.**

***Index Terms*—Fuel Level Sensor, Data Cleaning, Gaussian Mixture Model, Particle Swarm Optimization.**

I. Introduction

With the development of the Intelligent Transportation System (ITS) and Internet of Vehicle (IOV), vehicle monitor platform and the data plays an important role in the modern traffic system and is becoming easy to access through information and communication system designs with a collection of sensors, transmission of wireless vehicular networks and storage in remote databases. Vehicle data in this paper is a set of multi-dimensional data including speed, GPS coordinates, fuel level, steering angle and other real-time status data collected from vehicles. With the rising of fuel prices, business owners have paid increasing attention on these events during the driving process. On the other hand, with the increasing importance of energy conservation and environmental protection, accurate estimation of fuel consumption in vehicle data has become a decisive factor for traffic management and vehicle navigation, especially in routing optimization problems. To obtain an optimal route, not only travel time should be considered but also other factors, such as fuel consumption and emission. Y. Peng and X. Wang [1] proposed that the optimal vehicle routing schedule to minimize fuel consumption is probably different from the one to minimize travel distance. Ahn, Kyoungho, et al. [2] presented several hybrid regression models for estimating vehicle fuel consumption and emissions based on the instantaneous speed and acceleration levels of light-duty vehicles and light-duty trucks.

However, there are several types of error during the process of fuel level data acquisition. The reasons causing outliers in fuel level data can be divided into four kinds:

- Refueling: it will cause the fuel level to increase rapidly;
- Fuel spilling or gasoline theft: it will cause the fuel level to decrease rapidly;
- Errors of fuel level sensor or wireless vehicular network transmission: it will cause the fuel level to fluctuate anomalously;
- The large shake of vehicle: it will cause fuel level fluctuation.

Due to the fluctuation of vehicles and precision of fuel level sensors, there are many errors in collected original data by fuel level sensors. During the transmission and storage of fuel level data, stability of wireless vehicular networks and availability of data transformation for storage also cause errors in fuel level data. The quality of fuel level data is affected not only by the precision of fuel level sensors but also the quality of transmission networks. There are existing literature discussing the improvement of the precision of fuel level sensors [3] and the quality of transit networks [4]. Nevertheless, the challenges to cause errors or missing data remains when the fuel level data is collected by fuel level sensors, transmitted in wireless vehicular networks and stored in remote databases. Data cleaning is necessary before analyzing the data collected at the database [5]. In the process of cleaning and repairing the fuel level data, the former two types of abnormal data are caused by human behavior. The first one can be identified, captured and then remained after the process, while the second one should be identified and captured for human interpretation. The latter two are the noisy data, which should be cleaned and repaired during the process. To separate the former two from the latter two, threshold of average fuel consumption per minute and fuel charge are proposed in this paper.

This research was supported in part by the National Natural Science Foundation of China under Grant Nos. 61672082, U1564212 and 61711530247, Beijing Municipal Natural Science Foundation Nos. 4181002, Asa Briggs Visiting Fellowship from University of Sussex, Royal Society-Newton Mobility Grant (IE160920) and The Engineering, and Physical Sciences Research Council (EPSRC) (EP/P025862/1). *(Corresponding author: Xuting Duan)*

D. Tian Y. Zhu, X. Duan, J. Hu, J. Wang, Y. Wang are with Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China (e-mail: *duanxuting@buaa.edu.cn*).

Z. Sheng is with Department of Engineering and Design, the University of Sussex, Richmond 3A09, UK.

M. Chen is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China.

The acquisition of fuel level data is usually combined with other attributes of vehicle data, such as velocity and GPS coordinates. There are also many errors in these vehicle data and many literatures have paid attention to data processing of inaccurate velocity [6] and GPS coordinates data [7]. But there are very few studies of processing fuel level data after stored in remote databases for accurate fuel consumption data [8]. In fact, because high-precision fuel level sensor is very expensive, the fuel level sensors installed on most vehicles produce much noise. Hence there are some existing filtering method for fuel level data used in vehicle electronic system, e.g., wavelet transform [9], which is applied for direct signal from the fuel level sensor to estimate fuel consumption. But the source of its noise is less than the data in remote databases, such as the error from wireless communication. The sampling frequency is also much higher than the data in remote databases. Hence the existing filtering method is not suitable for the fuel level data at the background of IOV. To ensure the quality of fuel level data, a method for cleaning and repairing the fuel level data is proposed in this paper. It can identify the abnormal data by a modified Gaussian Mixture Model (GMM), and correct error data by a specific method based on linear interpolation.

The remainder of this paper is organized as follows. In Section II, the vehicle data related work is presented. The fuel data cleaning and repairing model is proposed in Section III. In Section IV, the effects of the model are discussed based on the smoothing spline and the comparative analysis. The concluding remarks are given in the final section.

II. Related Work

Vehicle data, which is collected by sensors, transmitted through wireless vehicular networks and stored in remote databases, is the fundamental for traffic management, establishment of traffic systems, and vehicle navigation. To acquire more accurate vehicle data, many scholars have studied methods to process and rectify noise of the original data in databases. S.B. Lazarus and I. Ashokaraj [10] proposed a vehicle localization method with sensors data fusion algorithm which combined the extended Kalman filter, interval analysis and covariance intersection. Y. Zhang $et\ al$ [11] adopted a Kalman filter arithmetic to improve vehicle GPS data accurace. Data fusion based on dead reckoning was proposed to deal with inaccurate vehicle position data due to the low accuracy of sensors and low calculating capability. Z. Zhang $et\ al$ [6] propose a method for cleaning incorrect probe vehicle data by a 3σ rule method and principal component reconstruction. Cubic spline interpolation method was used in [12] to correct the singular signal in the vehicle driving data such as vehicle speed, time, and position.

Although methods for accurate vehicle data such as position, velocity has widely studied, very few studies have investigated accuracy of vehicle fuel level data. With the increasing importance of energy conservation and environmental protection, the fuel consumption data has become a decisive factor for traffic management and vehicle navigation. H. Zhao $et\ al$ [8] proposed a method to analyze and process the outlier data of fuel consumption data. The threshold of detecting fuel consumption is known, and the situation of appearing continuous-time abnormal fuel level is not considered in this method. Some filtering method like Kalman filter, fast Fourier transform (FFT) and discrete wavelet transform (DWT) [13]have been widely used in denoising and outlier detection on sensor data. DWT has been developed rapidly in recent decades and has obvious advantage on temporal resolution [14]. But these filtering methods are applied for direct signal from the fuel level sensor, where the source of its noise is less than the data in remote databases and the sampling frequency is also much higher. Hence the existing filtering method is not suitable for the fuel level data at the background of IOV. In this paper, a model for cleaning and repairing fuel level data is proposed to acquire accurate fuel level data.

Outlier detection (also Anomaly detection) is an important part in this model. With the development of computer and communication technology, big data can be sensed, transmitted and stored for mining. But there are always some abnormal data during these processes. Data cleaning is aimed at ensuring data quality by discovering and correcting the errors or inconsistencies in data, including checking data consistency, dealing with invalid values and missing values and etc.. Outlier detection is a kind of method for discovering errors or inconsistencies in data. Statistical-based algorithm is used most widely and traditionally, which is generally suited to quantitative real data or at the very least quantitative ordinal data distributions where the ordinal data can be transformed to suitable numerical values for statistical processing. This limits their applicability and increases the processing time if complex data transformations are necessary before processing. Distance-based algorithm [15] has also been developed for different demands. But for unsupervised learning mission, Cluster-based algorithm [16] is better than traditional methods due to needless of data distribution and applied in outlier detection very successfully and extensively. A clustering based data mining algorithms was proposed by B. Liu $et\ al$ [17] to solve the outlier data problem. P. Zhang $et\ al$ [18] presented the Outlier Detection based on Cluster Analysis and Spatial Correlation(ODCASC) algorithm to acquire outliers. K-means clustering algorithm was adopted for outlier detection by dividing the data set into clusters in [19]. But these methods can not satisfy the demand in our case very well, because the accurate notion of outliers is different for different application domains [20], which means a method developed in one domain is usually not suitable for another.

The GMM is one of the cluster algorithms and it classifies data by describing them using multiple Gauss distributions. It is usually used with the Expectation Maximization (EM) algorithm to optimize the parameters of the GMM. However, EM algorithm has been proved to often fall into local optimum when optimizing the parameters of GMM, which will cause poor performance of GMM. In this paper, a modified GMM with the synchronous iteration method to optimize the parameters is proposed for outlier detection. Optimizing the parameters of GMM is actually an optimization problem and there are many artificial intelligence algorithms for solving this problem. A series of global search-based heuristic algorithms have been widely and successfully applied for optimization

problems, including the ant colony optimization algorithm (ACO) [21], genetic algorithm (GA) [22], Particle Swarm Optimization (PSO) algorithm [23] and etc.. PSO algorithm is one of the most successful and effective intelligence algorithms, variants of which and itself have been widely and successfully applied for optimization problems. PSO makes few assumptions about the problem being optimized and does not use the gradient of the problem, which means PSO does not require that the optimization problem be differentiable as is required by traditional optimization methods such as gradient descent. On the other hand, comparing with other heuristic algorithms like ACO and GA, PSO is of less complexity and quick constringency speed, but it's more likely to fall into local optimum. To eliminate the possibility of falling into local optimum when turning parameters of GMM by EM algorithm, the synchronous iteration method based on the PSO algorithm and the steepest descent algorithm is introduced in the modified GMM, and the comparison experiment shows the superiority of the proposed algorithm.

III. A FUEL LEVEL DATA CLEANING AND REPAIRING METHOD

In the vehicle monitor platform, there are about 30,000 vehicles and the data of vehicles is sent to database by GPRS (General Packet Radio Service) message of cellular network. These vehicles belong to different companies and it is difficult to detect the fuel level data for each vehicle by manual work. To complete this work, machine recognition must be introduced. However, the inaccurate data has a great influence on machine recognition. The fuel level data cleaning and repairing method is adopted to ensure the data quality for machine recognition. In the experiments, 50 vehicles' fuel level data collected in one month were used for testing the method. Since the data is updated per minute when the vehicle terminal is online which is not consistently, the sample serial-number is used as x axes of figures in this paper.

A. Pretreatment

For vehicles equipped with sensor and wireless communication units, fuel level data is transmitted with other vehicle properties data to the database with fixed time interval, such as every minute in this paper. The average fuel consumption data $avgfuel$ between two continuous time points of fuel level data can be calculated as:

$$avgfuel(t) = fl_t - fl_{t-1} \tag{1}$$

where fl_t is the fuel level data at time t, and fl_{t-1} is the fuel level data at the previous time interval. However, due to the wireless interference, limited fuel level sensors power and outlier from other malicious electric devices, there are a lot of missing values in the fuel level data, which leads to no continue data to calculate the $avgfuel$. The average fuel consumption data per minute of each vehicle is within a certain range and the abnormal average fuel consumption can be acquired through outlier detection. Before analyzing the average fuel consumption, it is necessary to get the accurate average fuel consumption data. What's more, the fuel value of

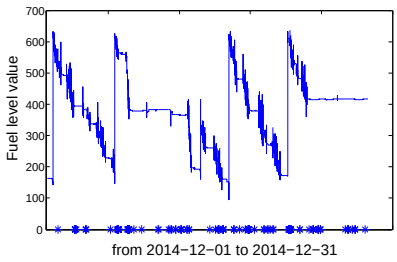


Fig. 1. Original fuel level data. The symbol * represents the missing data, the solid line represents the original fuel level data.

rising edges $RValue$ in the fuel level data is also calculated in this section to confirm the threshold of fuel charge. If there are some time points where positive $avgfuel$ are continuous, time point before the first time point will be considered as starting time st and the final time point will be considered as ending time et. If there are no other time points adjacent to a positive $avgfuel$, the time point at the positive $avgfuel$ will be considered as ending time et, and time point before the time points will be considered as starting time st. Then $RValue$ can be calculated as:

$$RValue(i) = {fl^{(i)}}_{et} - {fl^{(i)}}_{st} \tag{2}$$

where i means the ith rising edge in fuel level data; ${fl^{(i)}}_{et}, {fl^{(i)}}_{st}$ are respectively represent fuel level value at time point et, st of the ith rising edge.

The data collected on each vehicle over a period, such as one month, can be expressed as a matrix X in

$$X = \begin{bmatrix} T_1 & Lo_1 & La_1 & V_1 & Fuel_1 \\ T_2 & Lo_2 & La_2 & V_2 & Fuel_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ T_n & Lo_n & La_n & V_n & Fuel_n \end{bmatrix} \tag{3}$$

Where T_i represents the time point of data recording and Lo_i, La_i, V_i, $Fuel_i$ respectively means the longitude, latitude, velocity, and fuel level of the vehicle at time point T_i. n is the number of data collected in one month. To obtain the accurate fuel level data, the whole month should be divided into a lot of time quanta when the data is collected per minute. It is easy to be realized by time division, which means that when there are more than one minute between two consecutive time points in X, there should be two periods. Fig.1 shows the fuel data of one vehicle in one month. The fuel level of vehicle is valued as zero at the time of lacking data, and these data is treated as abnormal data in the outlier detection process. Then in continuous time quanta, $avgfuel$ can be calculated by the subtraction of two fuel level data at two continuous time points and $avgfuel$ at last time point is defined as zero to avoid affecting other data.

B. Outlier detection algorithm

To acquire the thresholds of normal fuel consumption and fuel charge, a modified GMM algorithm aimed at improving GMM-EM is proposed in this paper to calculate these thresholds. Gaussian mixture model and EM algorithm are firstly

proposed by Middleton [24]. Detailed and popular descriptions of the GMM and GMM-EM can be found in [25]. In GMM, the regularities of unknown distribution data X is assumed to be expressed by a Gaussian mixture model of k components, and the theoretical basis is the central limit theorem. In this paper, the maximizing log-likelihood function is used for calculating the parameters of Gaussian mixture model.

$$\max(L(X \mid (\pi_j; \mu_j; \sigma_j)) = \sum_{i=1}^{N} \log(\sum_{j=1}^{k} \pi_j g(x_i \mid (\mu_j; \sigma_j))) \tag{4}$$

which means that an unknown sample data $X = x_i, i = 1, 2, 3, ...$ can be expressed by a Gaussian mixture model with the parameters (π_k, μ_k, σ_k) obtained by maximizing the Log-likelihood function. $g(x_i, \mu_j, \sigma_j)$ is the j_{th} Gaussian distribution with mean value μ_j and variance σ_j. π_j is the possibility of x_i to choose the j_{th} component, which is the mixing coefficient.

In order to obtain the optimal maximum of Log-likelihood function, the EM algorithm is usually used to calculate the parameters of Gaussian mixture model, known as GMM-EM. However, the EM algorithm usually falls into local optimum and it may acquire local optimal parameters for GMM. So the modified GMM based on the synchronous iteration method to improve GMM is proposed in this paper. The reason for using EM algorithm to optimize parameters is that the Log-likelihood function is not easy to be derived and the classical mathematic methods cannot be used for it. However, the Log-likelihood function is easy to be derived by μ_j with known σ_j. The steepest descent algorithm is introduced to optimize the parameter μ. For σ_j, the Maximizing Log-likelihood function is actually an optimization problem with known μ_j. PSO algorithm is an effective intelligence algorithm, which is adopted to optimize the parameter σ_j in this paper. In each iteration of the modified GMM, the steepest descent algorithm and the PSO algorithm are synchronously iterative step. For another parameter π_j, which could be calculated as the probability of the x_i generated by components j:

$$\gamma(i, j) = \frac{\pi_j g(x_i \,|\mu_j, \sigma_j)}{\sum_{j=1}^{k} \pi_j g(x_i \,|\mu_j, \sigma_j)} \tag{5}$$

Then the number N_j of x_i generated by component j can be obtained by $\sum_{i=1}^{N} \gamma(i, j)$, and the coefficient $\pi_j = \sum_{i=1}^{N} \gamma(i, j)/N$, which is the means of the membership values over the N data points .

From the procedure of GMM with the synchronous iteration method, we can know that the iterations are actually establishing GMM and updating parameters (π, μ, σ). In fact, it is the same to the GMM-EM algorithm, but using a new way based on the PSO algorithm and the steepest descent algorithm to update parameters μ, σ separately instead of EM algorithm.

1) Calculating the optimal σ_j based on PSO: PSO algorithm is inspired by the predation behavior of birds. To acquire an optimal solution of optimization problems, each particle $x_{k,i}$ in the k iteration of PSO algorithm flies around the multidimensional search space by $v_{k,i}$. v_i and x_i is updated by Equ.(6) and Equ.(7).

$$v_{k+1,i} = \omega v_{k,i} + c_1 R_1 (x_{pbest,i} - x_{k,i}) + c_2 R_2 (x_{gbest} - x_{k,i}) \tag{6}$$

$$x_{k+1,i} = x_{k,i} + v_{k+1,i} \tag{7}$$

where ω, c_1, c_2 are the inertia weight and constriction factors of PSO algorithm and they are often valued as constant. R_1, R_2 are the random values in $[0, 1]$. $x_{k,i}$ means the position of the ith particle in k iteration and $v_{k,i}$ is the velocity. $x_{pbest,i}$ shows the present optimal position of the ith particle and x_{gbest} means the actual optimal position of the population. The update of $x_{pbest,i}$ and x_{gbest} depends on the objective function of PSO algorithm. The mechanism of updating is the core idea of PSO algorithm, which means that each particle exploring the multidimensional space is based on its own experience and the experience of the whole population. When iteration ends or the condition of acquiring the optimal value is established, the optimal x_{gbest} and the optimal value of the objective function will be actually found.

To optimize parameter σ, because PSO is usually used for minimum optimization problem, the negative value of Log-likelihood function Equ.(4) of GMM is defined as the objective function of PSO algorithm. Parameters σ_k is defined as the value of particle x_i and the interval of $(0, \infty)$ is considered as the multidimensional search space of PSO algorithm. For d-dimension X, σ_k is a D^2 dimension matrix. In order to confirm that the covariance matrix σ is a positive definite symmetric matrix in iterations, the method proposed in [26] is adopted in this paper. The cyclic Jacobi eigenvalue decomposition algorithm is used to parameterize the covariance matrix σ_k, so the D^2 dimension σ_k can be defined as $\sigma_k = V^\top \Lambda V$. Where Λ, a diagonal matrix $diag(\lambda_1, \lambda_2, ..., \lambda_d), where \lambda_i > 0$, composes eigenvalue of σ_k and V is constituted by corresponding unit feature vectors $(v_1, v_2, ..., v_d)$. V can also be calculated by QR decomposition, which means $V = QR$, where R is a unit matrix and Q is defined as

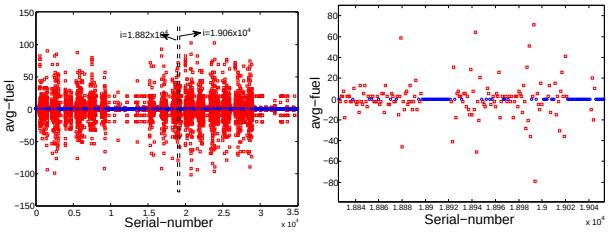
$$Q = \prod_{p=1}^{d-1} \prod_{q=p+1}^{d} G(p, q, \phi^{p,q}) \tag{8}$$

$G(p, q, \phi^{p,q})$ is Givens rotation matrix:

$$G(p, q, \phi^{p,q}) = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\phi^{p,q}) & \cdots & \sin(\phi^{p,q}) & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & & -\sin(\phi^{p,q}) & \cdots & \cos(\phi^{p,q}) & & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \tag{9}$$

The non-zero elements of Givens rotation matrix $G(p, q, \phi^{p,q})$ are:

$$\begin{cases} g_{k,k} = 1, if k \neq p, q \\ g_{p,p} = \cos(\phi^{p,q}) \\ g_{q,q} = \cos(\phi^{p,q}) \\ g_{p,q} = \sin(\phi^{p,q}) \\ g_{q,p} = -\sin(\phi^{p,q}) \end{cases} \tag{10}$$



(a) The whole cluster (b) The magnified portion of sub-figure(a)

Fig. 2. Fuel data $avgfuel$ cluster. The circles represent the normal fuel consumption data. The squares represent the abnormal fuel consumption data.

$g_{i,j}$ is the (i,j) value in $G(p,q,\phi^{p,q})$, and $\phi^{p,q}$ is in the interval $[-\pi/2, \pi/2]$. Therefore, a covariance matrix with $d(d-1)/2$ degree of freedom can be described with parameters $\left(\lambda_1, \lambda_2, ..., \lambda_d; \phi^{1,2}, ..., \phi^{p,q}, ..., \phi^{(d-1),d}\right)$. So the σ_j of jth component can be parameterized with $(\lambda_j^1, \lambda_j^2, ..., \lambda_j^d; \phi_j^{1,2}, ..., \phi_j^{p,q}, ..., \phi_j^{(d-1),d})$. After the optimal parameters $(\lambda_j^1, \lambda_j^2, ..., \lambda_j^d; \phi_j^{1,2}, ..., \phi_j^{p,q}, ..., \phi_j^{(d-1),d})$ is obtained by PSO algorithm, the optimal σ_j is established.

2) Calculating the optimal μ_j based on the steepest descent algorithm: For μ_j, the Log-likelihood function Equ.(4) can be derived by μ with parameters π_j and σ_j. Defining Δ as the $p(x_i \mid (\pi_j; \mu_j; \sigma_j)$, so Log-likelihood function is defined as:

$$L(X \mid (\pi_j; \mu_j; \sigma_j)) = \sum_{i=1}^{N} \log(\Delta) \tag{11}$$

Then the Log-likelihood function derived by μ is defined as:

$$\frac{\partial L}{\partial \mu_j} = \sum_{i=1}^{N} \frac{1}{\Delta} \frac{\partial \Delta}{\partial \mu_j} \tag{12}$$

where

$$\frac{\partial \Delta}{\partial \mu_j} = \pi_j \frac{1}{(2\pi)^{d/2} \mid \sigma_j \mid^{1/2}} \times \exp^{(-1/2)(x_i-\mu_j)^\top \sigma^{-1}(x_i-\mu_j)} \sigma^{-1}(x_i - \mu_j) \tag{13}$$

For derivable function optimal problems, there are many classical algorithms like Newton method, steepest descent, and conjugate gradient method. In this paper, steepest descent algorithm [27] is used for calculating the optimal μ of Log-likelihood function Equ.(4). Its iteration equation is as follows:

$$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)} \tag{14}$$

where $d^{(k)} = -\nabla f(x^{(k)})$ and λ_k is calculated as:

$$f(x^{(k)} + \lambda_k d^{(k)}) = min(f(x^{(k)} + \lambda_k d^{(k)})), \lambda_k >= 0 \tag{15}$$

When $\parallel d_k \parallel$ meets the condition $\parallel d_k \parallel < \varepsilon$ where $\varepsilon > 0$, the iteration is completed. In this way, the optimal μ can be obtained with the synchronous iteration of PSO algorithm for σ.

The procedure of the modified GMM algorithm is presented in Algorithm 1. For the outlier detection of $avgfuel$ and the cluster analysis of $RValue$, deciding the number k of cluster in the classification by the modified GMM is pre-requisite,

Algorithm 1 GMM with synchronous iteration method

procedure GMM PARAMETERS(optimal μ, σ)
  initialize c_1, c_2, ω and swarms(μ, σ)
  while do not meet the end of iteration conditions **do** ▷ get the optimal evaluation
    for each particle μ_j, σ_j of PSO algorithm **do**
      establish GMM;calculate π_j for each particle
      calculate the fitness function (negative Log-likelihood function) evaluation of each particle with μ_j, σ_j
      communicate with others; update $pbest_j, fpbest_j$ of each particle
    end for
    update $gbest, fgbest$ of whole swarms
    update v, σ of each particle
    for each particle μ_j, σ_j of PSO algorithm **do**
      calculate $d_j = -\nabla f(\mu_j, \sigma_j)$ where f is the negative Log-likelihood function.
      calculate λ by equation (15)
      calculate $\mu_j = \mu_j + \lambda d_j$
    end for
  end while
  return μ, σ ▷ The optimal evaluation is μ, σ
end procedure

which is the most important step. The defined $avgfuel(t)$ which actually is the fuel level value at time t minus the fuel level value at time $t-1$, it can be called the fuel consumption per minute, but does not mean the real value. Actually, data $avgfuel(t)$ is $realfuel(t)$ plus $noise(t)$, represented by the following equation:

$$avgfuel(t) = realfuel(t) + noise(t) \tag{16}$$

where $realfuel(t)$ represents the real fuel consumption per minute and $noise(t)$ signifies the fluctuation. Because the vehicle jolts continually when it's running, the fuel level data fluctuate frequently, leading to the difficulty for the fuel level sensors to acquire accurate fuel level value, which means the defined avgfuel(t) include both many positive and negative values around zero. In fact, the fluctuation from the movement of vehicle itself is very small, and it can be treated as normal data. But large shake can cause large fluctuations of fuel level data. Whats more, deviation of network transmission, deviation of storage, refueling, and fuel spilling will cause large fluctuations of $avgfuel$. In one word, the absolute value of $avgfuel$ can be divided into two main categories including normal little fluctuations and abnormal large fluctuations, hence the number k of clusters is set as 2 in the modified GMM, and one sample result are presented in Fig.2.

In Fig.2, the data marked as circles is the normal data which means $realfuel$ combining $noise$ caused by movement of the vehicle itself, and the data marked as squares is outlier of the data $avgfuel$. The threshold 2.5 of normal data can also be obtained in this way. It means that if $avgfuel(t)$ is less than -2.5 or more than 2.5, the fuel consumption $avgfuel(t)$ will be considered as abnormal fuel consumption. The time points of refueling, fuel spilling or gasoline theft are included

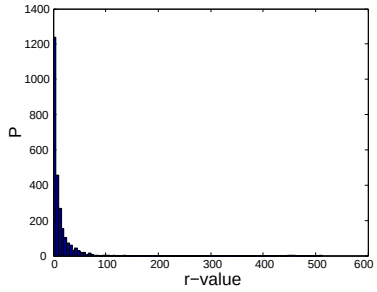


Fig. 3. Histogram of posterior probability distribution for $RValue$

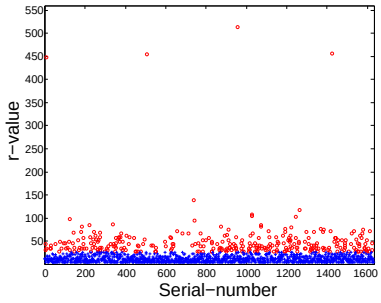


Fig. 4. Fuel data $RValue$ cluster. The circles show there are may be some rising edges of being refueled. The pluses represent the noise.

in the abnormal data $abfuel$. These detailed description of this threshold is given in subsection C.

For cluster analysis of $RValue$, the histogram of the posterior probability distribution for $RValue$ is shown in Fig.3. It can be found that there are many rising edges in the month data and $RValue$ is almost zero. To reduce the influence of normal fuel consumption, $RValue$, which is less than the threshold of normal fuel consumption, is cleaned away before cluster analysis. In fact, $RValue$ may be caused by the fluctuation of vehicles or the deviation of network transmission, which means $RValue$ is uncertain and small. For refueling, the $RValue$ will be large because the vehicle is usually refueled when it is lack of fuel. Therefore, to acquire the threshold of fuel charge and remove the influence of the fluctuation of vehicle, the number k of clusters in the modified GMM for the cluster analysis of $RValue$ is also set as 2. The result of clusters is shown in Fig.4. The threshold value of fuel charge is 26.3, which means that if the $RValue$ of a rising edge is higher than the threshold, there may be a time point of refueling, and if not, there will not be.

C. Classifying abnormal time quantum

If $avgfuel(t)$ is abnormal and $avgfuel(t-1)$ is normal judged by the threshold of normal fuel consumption, the judgement of $avgfuel(t+1)$ based on the fuel level data at time t is inaccurate, whatever the fuel level data at time $t+1$ will be, due to the inaccurate fuel level data at time t. As the results shown in Fig.5, there are continuously abnormal data through vision at the time points of the sixth and seventh symbol *. But if only outlier detection of fuel level is used, the latter will be considered as normal data. To solve this problem, the method of identifying data by time quantum instead of time point is proposed. It means that the abnormal time quantum, in which more than one abnormal fuel consumptions are detected by the GMM with the synchronous iteration method, will be obtained in this way.

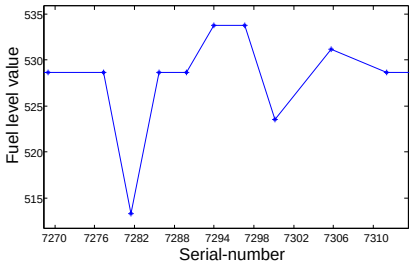


Fig. 5. Continuously abnormal data. The stars represent the fuel level value at these time points.

In the pretreatment stage we can get the continuous time quantum, and the abnormal data $abfuel$ through outlier detection of fuel level. So the time length of neighbouring data $abfuel$ in the same continuous time quantum can be calculated as:

$$timelength(t) = T(abfuel(t)) - T(abfuel(t-1)) \quad (17)$$

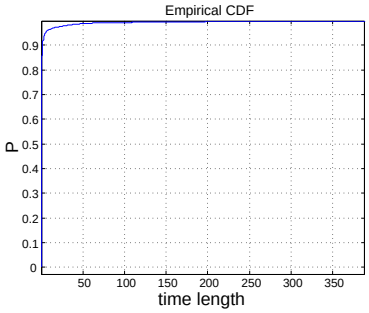
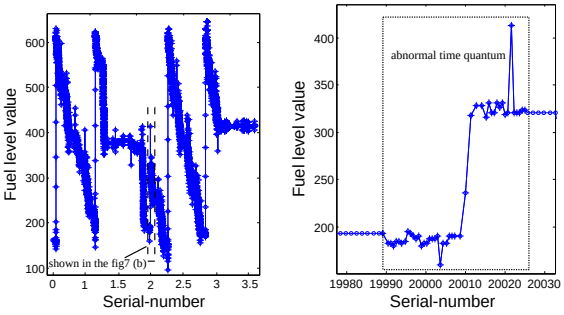


Fig. 6. Cumulative distribution formula of time length.

The probability distribution of time length $p(timelength)$ is proposed to detect the data $avgfuel$ between neighbouring $abfuels$. One sample of the probability distribution of time length is shown in Fig.6. It can be found that almost all the data $timelength$ is small, which means that there are just several data $avgfuel$ in most time quanta, thus making the influence of abnormal data prominent. It is also difficult to confirm that whether the data in a short $timelength$ is normal. In fact, there are many data $timelength$ which are equal to 1, meaning many abnormal data $abfuel$ is adjacent. In the adjacent abnormal data $abfuel$, the latter is influenced by the former and there is actually just an abnormal data. In Fig.6, the threshold $ThrTime = 5$ of data $timelength$ is defined as the value of $timelength$ when $p(timelength) = 0.95$. If a data $timelenth(i)$ is smaller than $ThrTime$, the data $avgfuel$ from the former $abfuel$ to the latter $abfuel$ will be considered as a time quantum, and the time of the former data $abfuel$ is considered as the starting point while the time of the latter data $abfuel$ is considered as the ending point. The



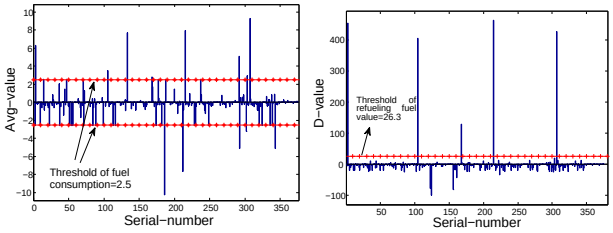
(a) Abnormal data in one month (b) The magnified portion

Fig. 7. Abnormal time quantum. The line with pluses represent the abnormal time quantum. The line with circles represent the normal time quantum.

way to judge whether there is a time point of refueling, fuel spilling or gasoline theft is as follows: firstly, the $Formeravg$ is calculated by the average of five data before the starting point; secondly, calculate $Latteravg$ by the average of the five data after the ending point; and finally, compare the Avg-value which is equal to $(Latteravg - Formeravg)/timelength$ with fuel threshold of normal fuel level data, and compare the D-value which is equal to $Latteravg - Formeravg$ with the threshold of fuel charge. Using the Avg-value and D-value to judge whether the current abnormal time quantum data is in accordance with the specified threshold of abnormal fuel consumption data. If the Avg-value does not belong to the fuel threshold of normal data, there will be a time point of refueling, fuel spilling or gasoline theft.

One example time quantum is shown in Fig.7. Because there are too much data displayed in one picture, it may show that most data is abnormal. When the time length of data is shrunk to one week or one day, it can be shown very clearly. The magnified portion of Fig.7a is shown in Fig.7b. Time quantum of refueling means that Avg-value is bigger than the maximum of normal fuel level data acquired in the outlier detection of fuel level and D-value is bigger than the threshold of fuel charge. Fuel spilling or gasoline theft time quantum means that Avg-value is smaller than the minimum of normal data. As there is no threshold for identifying the fuel spilling or gasoline theft, it is difficult to distinguish that whether there is fuel spilling or gasoline theft in the time quantum when Avg-value is smaller than the minimum of normal data.Amidan2005Data So the time quantum is used to confirm whether there are time points of fuel spilling or gasoline theft instead of calculating the fuel level falling edge.

One example of the Avg-value and the D-value is shown in Fig.8. The solid line with stars in Fig.8a represents the threshold of normal fuel consumption acquired by outlier detection, and the threshold of fuel charge is identified as the solid line with pluses in Fig.8b. Five time quanta of refueling can be captured by comparing Avg-value with the threshold of normal fuel consumption data and by comparing D-value with the threshold of refueling fuel value. There are also five time quanta of refueling in the fuel level data. It means that the time points of refueling are all in the time quantum of refueling acquired by the model. Whats more, the time quanta



(a) Identify the Avg-value (b) Identify the D-value

Fig. 8. Compare Avg-value and D-value with their thresholds.

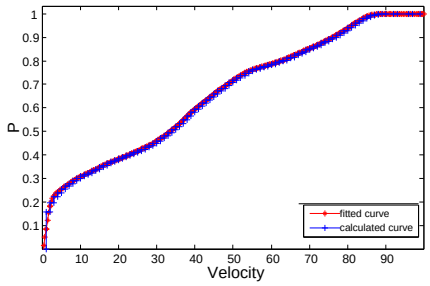


Fig. 9. The cumulative probability distribution of velocity. The solid line with stars represents the fitted curve. The solid line with pluses represents the calculated curve.

of fuel spilling or gasoline theft are also obtained by the same method presented in Fig.8a. It is not accurate enough to clean fuel level data in the abnormal time quantum when there are other outliers. Therefore, the way to acquire accurate time points of being refueled is proposed through velocity analysis because refueling must be completed in the stopping state.

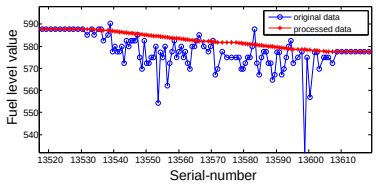
D. Velocity analysis for accurate time points of being refueled

From the above presentation, the abnormal time quantum with time points of refueling is acquired, but it is not accurate enough. Through the data acquired in the above sections, there are perhaps many large fluctuations and it is not easy to determine the time points of refueling in the abnormal time quantum. Like in Fig.8b, only knowing the fuel level data is not enough to determine the accurate time points of refueling because of the influence of other $abfuel$. Another property of vehicle is introduced to solve this problem that vehicle must be stopped when it's refueled. Due to the deviation of network transmission, velocity equal to zero does not mean that the vehicle is really stopping. So the threshold of velocity for confirming that velocity less than or equal to threshold means the vehicle should be stopped. But before the velocity analysis, all the values of fuel level rising edges in time quanta of refueling should be calculated to compare with the D-value. If there is just one value of fuel level rising edge, the fuel level rising edge is the time point of refueling. If there are more than one values of fuel level rising edge meeting the condition, these fuel level edges should be judged by velocity analysis.

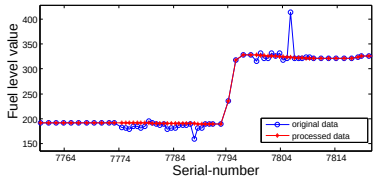
Choosing the velocity data which is unequal to zero from data X, and the cumulative probability distribution of velocity is shown in Fig.9. The solid line with pluses presents the cumulative probability distribution of velocity. It is discrete

and not enough to obtain the exact velocity threshold. To acquire the threshold of velocity and confirm that whether the vehicle is stopping, the fitting curve of the cumulative probability distribution is proposed and it is expressed as a solid line with stars in Fig.9. The confidence coefficient of these velocity data is determined as 0.05, and the confidence interval can be calculated by the fitting curve of the cumulative probability distribution. The confidence interval of whether the vehicle is in the stopping state is $[0, 0.2771]$. It means that if the velocity is less than 0.2771, the vehicle is regarded as in the stopping state. When the vehicle is being refueled, it must be stopped, which means that the velocity at these time points should be lower than the velocity threshold.

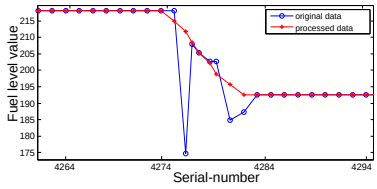
E. The abnormal time quantum



(a) Abnormal time quantum of noise



(b) Abnormal time quantum of being refueled



(c) Abnormal time quantum of fuel spilling or gasoline theft

Fig. 10. Process of the abnormal time quantum. the solid line with stars represents the processed fuel level data. the solid line with circles represents the original fuel level data.

From the above presentation, the abnormal time points acquired in outlier detection is defined as abnormal time quantum. In other words, the data in abnormal time quanta is entirely identified as abnormal data regardless of whether the data is abnormally acquired in the outlier detection stage. The data in abnormal time quanta is all cleaned and repaired. First, if the Ava-value is lower than the threshold of fuel consumption acquired in the outlier detection stage, it means that there are no time points of refueling, fuel spilling and gasoline theft in the quantum. These abnormal time quanta are of the third and the forth type, such as large shake leads to inaccurate fuel level data. To repair these data, the continuous time points, of which the velocity is less than the velocity threshold obtained in velocity analysis, should be took out from the abnormal time quantum. If there are some parking time points, the data exclusive of these parking time points in the abnormal time quantum will be valued by linear interpolation, and the fuel level of these parking time points will be valued as the average of fuel level values between the time points before and after these parking time points. If these velocities in the abnormal quantum are entirely higher than the velocity threshold, all the fuel level data will be repaired by linear interpolation. One example of the noise abnormal time quantum is shown in Fig.10a, where the solid line with circles shows the original data and the solid line with stars means the fuel level data after processing.

Second, to process the abnormal time quantum, in which there are time points of refueling, it is necessary to take out the time points of refueling from the abnormal time series, because if the time points of refueling do not be taken out, the data repaired by linear interpolation in the whole abnormal time quantum will be a linear rising edge, thus making large errors. In fact, when the vehicle is being refueled, it must be in stopping state and the fuel level will be rising rapidly. When it is not being refueled, the vehicle is using the fuel and the fuel level is declining. So the abnormal time quantum is seen as two abnormal time quanta divided by the time points of refueling, and these two abnormal time quanta are processed in the same way as the first one. The data at the time points of refueling will be remained after the process. One example of the refueling time quantum is shown in Fig.10b.

Finally, the abnormal time quantum of fuel spilling or gasoline theft is identified because there is no standard to define fuel spilling and gasoline theft, the abnormal time quantum is actually the time quantum of abnormal fuel consumption. But after processing, the results of this type of abnormal time quantum after the whole model processing are identified. One example of the fuel spilling or gasoline theft time quantum is shown in Fig.10c.

By now, the main parts of the method have been presented. The outlier detection and cluster analysis are based on the modified GMM with the synchronous iteration method, which can be used to determine the threshold of average fuel consumption. And the time length and velocity analysis scheme can help to identify four kinds of outlier fuel level data more accurately. The whole process of the method is shown in Fig.11.

IV. Experiments and Results

A. Parameters setting

In the experiments, all of the parameters are shown in the TABLE I. Parameters of PSO algorithm in modified GMM are set in consideration of applicability according to [28], where $c1 = c2 = 1.49618$ and $w = 0.729844$. Considering the real data, the dimension, population size and fitness function are valued as 2, 30 and the log-likelihood function of GMM. As described in section III, other parameters in the process of clearing fuel level data are kinds of $avgfuel$, kinds of $RValue$, $Probabilityoftimelength$ and $Confidencecoefficientofvelocity$, which are respectively valued as 2, 2, 0.95 and 0.05.

Fig. 11. The process of the cleaning and repairing method. (FC: fuel consumption; T: threshold; ATQ: abnormal time quantum.)

TABLE I
PARAMETERS SETTING

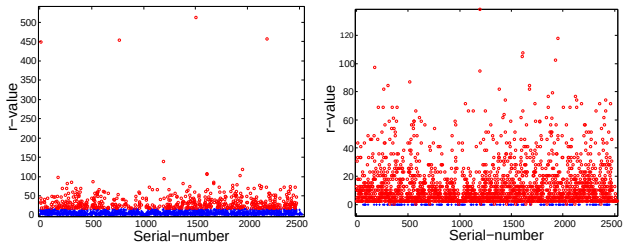
Indicators	Value
Population Size	50
Dimension	2
Parameters of PSO	c1=c2=1.49618,w=0.729844
Fitness function	Log-likelihood function of GMM
Kinds of avgfuel	2
Kinds of RValue	2
Probability of time length	0.95
Confidence coefficient of velocity	0.05

B. Comparison between modified GMM and GMM-EM

The aim of modifying GMM is to eliminate the possibility of falling into local optimum when turning parameters of GMM by EM algorithm. GMM is used in this paper for outlier detection, means acquiring threshold for detection. So stability of experiment results is an important indicator for evaluate these two algorithms. What's more, according to the theory of Gaussian mixture model, the value of Log-likelihood function is another important factor to evaluate the results of GMM, and the value of Log-likelihood function is larger which means that the results of GMM are better. In the experiments, the same fuel rising edges data were sent to the modified GMM proposed in this paper and the GMM-EM, and each of them run 50 times respectively. The results are shown in TABLE II, Fig.4 and Fig.12.

TABLE II
COMPARISON BETWEEN MODIFIED GMM AND GMM-EM ALGORITHM

Indicators	GMM-EM/times	modified GMM/times
AvgL	-9.7235e+03	-9.3267e+03
MaxL	-9.7010e+03/45	-8.4680e+03/8
MinL	-9.9263e+03/5	-9.5040e+03/42



(a) L=MaxL (b) L=MinL

Fig. 12. Fuel data $RValue$ cluster based on GMM-EM. The circles show there are may be some rising edges of being refueled. The pluses represent the noise.

In TABLE II, AvgL, MaxL and MinL respectively represents average, maximum and minimum value of Log-likelihood function. Times means frequency of acquiring almost same value of Log-likelihood. From TABLE II, it can be found interestingly that only two kinds of Log-likelihood value can be obtained respectively by GMM-EM and modified GMM. Fig.4 shows the result of clusters based on modified GMM and Fig.12 is based on GMM-EM. From these two figures, modified GMM can steadily classify $RValue$ and acquire threshold for outlier detection. While for GMM-EM, only when obtaining maximum value of Log-likelihood in Fig.12a, it is applicable for outlier detection, and times of minimum value in Fig.12b are inapplicable, which can be considered as falling into local optimum. So the possibility of falling into local optimum in EM algorithm is well eliminated by using PSO and steepest descent algorithm. What's more, the average, maximum, and minimum value of the Log-likelihood function based on the modified GMM are all higher than the ones of GMM-EM, which means that the modified GMM is distinctly superior to GMM-EM on the results of the value of Log-likelihood function. In a word, The experimental results

TABLE III
PROBABILITIES CALCULATED BY ORIGINAL DATA, GMM-EM AND MODIFIED GMM

value	original data	GMM-EM	modified GMM
2	0.2093	0.0450	0.0518
4	0.0621	0.0583	0.0700
6	0.0289	0.0644	0.0727
8	0.0119	0.0607	0.0582
10	0.0103	0.0489	0.0368
12	0.0067	0.0338	0.0195
14	0.0051	0.0203	0.0104
16	0.0016	0.0109	0.0071
18	0	0.0057	0.0065
20	3.9573e-04	0.0033	0.0066

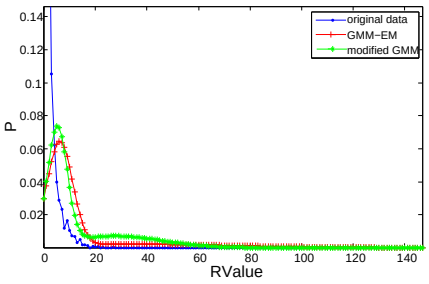
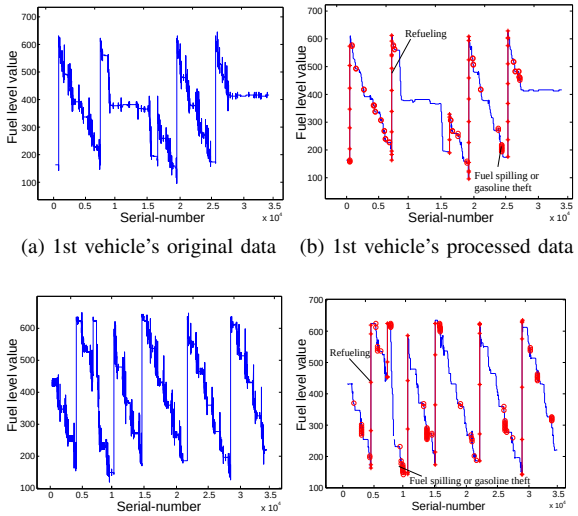


Fig. 13. Probability density distribution. The solid line with points represents the original data. the solid line with stars represents modified GMM and the solid line with plus represents the GMM-EM.

TABLE IV
PARAMETERS OF TWO VEHICLES IN THIS MODEL

parameters	first vehicle	second vehicle
threshold of normal fuel consumption	2.5	2.4
threshold of refueling fuel value	26.3	30.2
ThrTime	5	6
threshold of velocity	0.2271	2.4215



(a) 1st vehicle's original data (b) 1st vehicle's processed data

(c) 2nd vehicle's original data (d) 2nd vehicle's processed data

Fig. 14. Comparison between processed fuel level data and original data.

prove the modified GMM is more applicable for the model of clearing and repairing fuel level data.

On the other hand, the aim of GMM is using multiple Gauss distributions to express the data with an unknown distribution. In this way, as most rising value data is in the interval [0,20], we choose 10 values $(2, 4, ..., 20)$ of the data and calculate the probability of these values by the value frequency. The parameters of GMM-EM and the modified GMM are the optimal parameters acquired from the above experiments. The probability distribution curves are shown in Fig.13 separately, and the probabilities of these values are shown in TABLE III. It can be found that most of the probability density distribution curves established by the modified GMM are nearer to the original data density distribution curve. And the experimental results shown in TABLE III also show that most density values acquired by the modified GMM are nearer to the density value of the original data.

C. The Cleaning and Repairing Performance

To test the ability of the fuel level data cleaning and repairing method, the data of two vehicles collected over one month is used. After the processes of the outlier detection of $avgfuel$, cluster analysis of $RValue$, and the probability analysis of time length and velocity, the parameters calculated for each vehicle are shown in TABLE IV, which have been introduced in detail in section III. The calculated thresholds are different according to different data, but the causes of abnormal data, the method for acquiring abnormal time quantum, the threshold for classifying the abnormal time quantum, and the repairing process are all the same. For instance, the threshold of refueling fuel value of each vehicle are approximate and successful in identifying the time quantum of refueling.

Two vehicles' fuel level data of one month are shown in Fig.14. It can be found that most of the fluctuations of fuel level data caused by large shake, fuel level sensors or wireless network transmissions errors have been cleaned and repaired. Meanwhile the time quantum of refueling and fuel spilling or gasoline theft have been identified and captured. The solid line with stars shows the time points of refueling, and the solid line with circles shows that there may be some time points of fuel spilling or gasoline theft. Compared to the original data curve in Fig.14a and Fig.14c, the time points of refueling are well identified and captured. There are some errors to fuel spilling or gasoline theft. In fact, the fuel level data through the model is also not accurate enough, and it needs the following further process to get more accurate value.

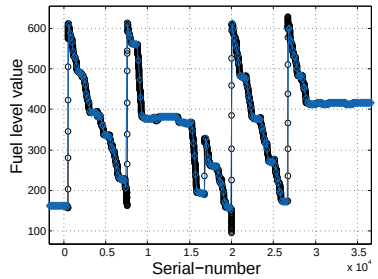
Because the real fuel level data is not prescient, many indicators of mathematical statistics cannot be used for identifying this model and smoothing spline is a method of curve fitting using spline function, which is widely used in data prediction and estimation [29]. In order to evaluate the effects of the proposed method, a comparison experiment with the widely used filtering method based on discrete wavelet transform (DWT) is conducted. The wavelet scaling function is Daubechies4, the decomposition level is 5, the filtering method is based on wavelet shrinkage and the parameter of filter is well adjusted to get a good result [30]. In this way the original data and processed data are fitted by the smoothing spline, and then the numerical results of these fitting data can be used to prove the effectiveness of our method. In the smoothing spline results, there are several indicators to evaluate the performance of fitting.

The sum of squares due to error (SSE):

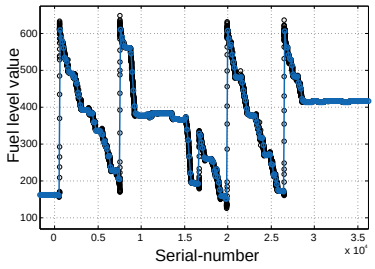
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{18}$$

Root mean squared error (RMSE) :

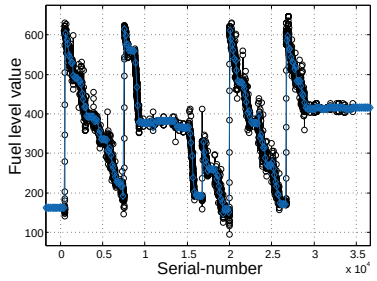
$$RMSE = \sqrt{SSE/n} = \sqrt{\frac{1}{dfe} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{19}$$



(a) The proposed method



(b) The wavelet transform



(c) The original data

Fig. 15. The fitting curve of original and processed fuel level data. The solid line with diamonds shows the fitted fuel level data. The solid line with circles shows the original fuel level data.

where $dfe = n - opp - 1$, and opp is the order of polynomial in smoothing spline.
Coefficient of determination (RSquare) :

$$RSquare = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{20}$$

Degree-of-freedom adjusted coefficient of determination (AR-Square) :

$$ARSquare = 1 - \frac{\sqrt{\frac{1}{dfe}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}} \tag{21}$$

In equations (18),(19),(20),(21), y_i is the value of the original data and $\hat{y}_i$ is the value of the fitted value. n is the number of the data and $\bar{y}_i$ is the mean value of the data. If SSE and RMSE are closer to 0, which means that the model selection and fitting is better, and the data prediction is more successful. The value of RSquare and ARSquare are in the interval [0,1] and the closer value are to 1, the better the fitting effect is. The fitting curve of the processed data by the model and the original data are shown in Fig.15, and the indicators are represented in the TABLE V. It can be found that all the indicators of smoothing spline declare that the performance of the proposed method is better than DWT, and more better than the original data. The value of SSE and RMSE based on the processed data by proposed method are extremely smaller than those of the original data. The RSquare and ARSquare also means a good fitting of the processed data by our method. In fact, although the smoothing effect of DWT is close to the proposed method, the wavelet-based denoising method drop out high-frequency information roughly. The noise type distinguished by DWT is also limited, such as gaussian white noise. Specially, the time quantum of refueling, fuel spilling or gasoline theft can be identified in our method, which means the processed data by our method which retain more valuable information than the classical filtering method.

TABLE V
INDICATORS OF SMOOTHING SPLINE

Indicators	proposed method	Wavelet Filter	original data
SSE	1497	5006	2.373e+05
RMSE	0.3264	0.7652	4.111
RSquare	1	1	0.9995
ARSquare	1	1	0.9987

V. CONCLUSION

At present, very few research studied the errors in the fuel level data of vehicles, which causes the inaccurate fuel consumption data of vehicles. In this paper, a model for fuel level data cleaning and repairing is proposed to ensure the quality of data. The modified GMM algorithm is proposed to detect fuel consumption and analyze the fuel value of rising edges. Furthermore, different from time points methods, the time quantum is proposed to evaluate abnormal fuel consumption. Finally, the effects of this model has been shown in the comparison experiments between the original fitting data and the processed fitting data using the proposed repairing method and DWT by smoothing spline. Meanwhile, comparison experiments between modified GMM and GMM-EM also shows the applicability and availability of the modified GMM. The results of experiments have shown that the original fuel level data with noise is well cleaned through the model and the time quantum of refueling, fuel spilling or gasoline theft is identified, which retain more valuable information than the classical filtering method. In our future research, we would like to extend the proposed model with consideration of the big data architecture and real-time data process method.

REFERENCES

[1] Y. Peng and X. Wang, "Research on a vehicle routing schedule to reduce fuel consumption," in *Measuring Technology and Mechatronics Automation, 2009. ICMTMA '09. International Conference on*, 2009, pp. 825–827.

[2] K. Ahn, H. Rakha, A. Trani, and M. Van Aerde, "Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels," *Journal of Transportation Engineering*, vol. 128, no. 2, pp. 182–190, 2013.

[3] C. Wang, "A capacitive sensor based fuel level measurement approach for a lorry," *Shandong Science*, vol. 25, pp. 83–86, 2012.

[4] W. Jiang, X. Chen, and Z. Zhong, "Analysis and optimization of gsm-r qos impacts on ctcs-3 onboard-trackside data transition," *Journal of Beijing Jiaotong University*, vol. 35, no. 5, pp. 17–20, 2011.

[5] L. Wang, L. D. Xu, Z. Bi, and Y. Xu, "Data cleaning for rfid and wsn integration," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 408–418, 2013.

[6] Z. Zhang, D. Yang, T. Zhang, Q. He, and X. Lian, "A study on the method for cleaning and repairing the probe vehicle data," *Intelligent Transportation Systems IEEE Transactions on*, vol. 14, no. 1, pp. 419–427, 2013.

[7] E. I. Laftchiev, C. M. Lagoa, and S. N. Brennan, "Vehicle localization using in-vehicle pitch data and dynamical models," *Intelligent Transportation Systems IEEE Transactions on*, vol. 16, no. 1, pp. 206–220, 2015.

[8] H. Zhao, J. Wang, and R. Zhao, "The outlier analysis and process of fuel consumption data based on .net," *MECHATRONICS*, vol. 16, no. 11, pp. 68–71, 2010.

[9] S. Shin, "Industrial application of wavelet analysis," in *International Conference on Wavelet Analysis and Pattern Recognition*, 2008, pp. 607–610.

[10] S. B. Lazarus, I. Ashokaraj, A. Tsourdos, R. Zbikowski, P. M. G. Silson, N. Aouf, and B. White, "Vehicle localization using sensors data fusion via integration of covariance intersection and interval analysis," *Sensors Journal IEEE*, vol. 7, no. 9, pp. 1302–1314, 2007.

[11] Y. Zhang, J. Shen, and C. Hu, "The research on vehicle gps position based on kalman filter," *Computer Knowledge & Technology*, 2009.

[12] H. Chen, H. Liu, S. Qiao, and Y. Wang, "Analysis of vehicle driving data based on cubic spline interpolation," *Automobile Technology*, 2013.

[13] H. Musoff and P. Zarchan, "Fundamentals of kalman filtering: A practical approach, second edition," *Progress in Astronautics and Aeronautics*, vol. 190, no. 8, p. 83, 2015.

[14] Z. Qin, L. Chen, and X. Bao, "Wavelet denoising method for improving detection performance of distributed vibration sensor," *IEEE Photonics Technology Letters*, vol. 24, no. 7, pp. 542–544, 2012.

[15] M. Radovanovi, A. Nanopoulos, and M. Ivanovi, "Reverse nearest neighbors in unsupervised distance-based outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1369–1382, 2015.

[16] D. Lian, L. Xu, L. Ying, and J. Lee, "Cluster-based outlier detection," *Annals of Operations Research*, vol. 168, no. 1, pp. 151–168, 2009.

[17] B. Liu, G. Xu, Q. Xu, and N. Zhang, "Outlier detection data mining of tax based on cluster," *Physics Procedia*, vol. 33, pp. 1689–1694, 2012.

[18] P. Zhang, X. Feng, and J. G. Zhou, "Outlier detection technique based on cluster analysis and spatial correlation in wireless sensor networks," *Application Research of Computers*, vol. 30, no. 5, pp. 1370–1364, 2013.

[19] R. Pamula, J. Deka, and S. Nandi, "An outlier detection method based on clustering," in *Emerging Applications of Information Technology, International Conference on*, 2011, pp. 253–256.

[20] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *Acm Computing Surveys*, vol. 41, no. 3, pp. 75–79, 2009.

[21] Y. Chen, D. Miao, and R. Wang, "A rough set approach to feature selection based on ant colony optimization," *Pattern Recognition Letters*, vol. 31, no. 3, pp. 226–233, 2010.

[22] R. H. Sheikh, M. M. Raghuwanshi, and A. N. Jaiswal, "Genetic algorithm based clustering: A survey," in *International Conference on Emerging Trends in Engineering and Technology*, 2008, pp. 314–319.

[23] A. A. A. Esmin, R. A. Coelho, and S. Matwin, "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data," *Artificial Intelligence Review*, vol. 44, no. 1, pp. 23–45, 2015.

[24] D. Middleton, "Statistical-physical models of urban radio-noise environments - part i: Foundations," *Electromagnetic Compatibility, IEEE Transactions on*, vol. emc-14, no. 2, pp. 38–56, 1972.

[25] Y. M. Omar, A. A. Ghaferi, and M. Chiesa, "What is the expectation maximization algorithm?" *Applied Physics Letters*, vol. 26, no. 8, pp. 897–899, 2015.

[26] C. Ari and S. Aksoy, "Maximum likelihood estimation of gaussian mixture models using particle swarm optimization," in *2010 International Conference on Pattern Recognition*, 2010, pp. 746–749.

[28] G. Venter and J. Sobieszczanskisobieski, "Particle swarm optimization," in *International Conference on Biomedical Engineering and Informatics*, 2012, pp. 129–132.

[29] N. E. Helwig and P. Ma, "Fast and stable multiple smoothing parameter selection in smoothing spline analysis of variance models with large samples," *Journal of Computational and Graphical Statistics*, vol. 24, no. 3, pp. 715–732, 2015.

[30] D. Pastor and A. M. Atto, *Wavelet Shrinkage: From Sparsity and Robust Testing to Smooth Adaptation*. Birkhuser Boston, 2010.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.